⋆HIERARCHICAL BAYESIAN MODELING WITH ABC AND IMPORTANCE SAMPLING⋆

BAYESIAN COMPUTING FOR ASTRONOMICAL DATA ANALYSIS

JESSI CISEWSKI

To use R at the RCC:

```
ssh -X lionxv.rcc.psu.edu
module load R
R
```

Open

```
abs_is_lab.R
```

in your favorite text editor.

For each problem, first copy and paste the relevant sections of code so it performs the needed calculations for the default parameter values. You'll be asked to make the changes to the R code. I suggested editing the source in a local text editor and copying and pasting the revised code into the R window.

1. Problem 1: Binomial Example

The data are a sample of 1's and 0's coming from $Y_i \sim \text{Bernoulli}(p)$ where n = sample size, $p = P(Y = 1)$. If $Y = \sum_{i=1}^{n} Y_i$, the $Y \sim \text{Binomial}(n, p)$ with a likelihood defined as

$$L(p \mid y) = \binom{n}{y} p^y (1-p)^{n-y},$$

where $y = \sum_{i=1}^{n} y_i$. We are going to pretend we do not know the likelihood function, and consider the following HBM for this problem:

$$Y \sim \text{Binomial}(n, p)$$
$$p \sim \text{Uniform}(0, 1)$$

Do the following:

(a) The sample size, `n`, is currently set to 20. Try running the code with different values for `n` (e.g. `n = 10, 50, 100`). What happens to the true posterior as `n` increases? What happens as `n` decreases?

(b) Set `n = 20`. The *particle* sample size, `N`, is currently set to 100. Try running the code with different values for `N` (e.g. `N = 25, 200, 500`). What happens to the ABC posterior as `N` increases? What happens as `N` decreases?

(c) Set `N = 200`. The tolerance, `epsilon`, is currently set to 0. Try running the code with different values for `epsilon` (e.g. `epsilon = 0.01, 0.1, 1`). What conclusions can you draw?

(d) The probability, `p`, is currently set to 0.75. Try running the code with different values for `p` (between 0 and 1 since it is a probability). What happens to the true posterior as you change this?

(e) The distance function, `rho`, is currently set to be $\rho(y, x) = |\sum_{i=1}^n y_i - \sum_{i=1}^n x_i|/n$. Try running the code without the absolute value in the distance function (i.e. $\rho(y, x) = (\sum_{i=1}^n y_i - \sum_{i=1}^n x_i)/n$). What happens to the ABC-posterior? Why does this happen? *Hint: consider the values of the means of the generated data (i.e., the generated* $\hat{p} = n^{-1}\sum_{i=1}^n x_i$*) that are going to be accepted with this distance function.*

2. Problem 2: Gaussian Example

Suppose we have a sample $i = 1, \ldots, n$ of $Y_i \sim N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. Consider the following HBM:

$$\begin{aligned} \mu &\sim N(\mu_0, \sigma_0^2) \\ Y_i \mid \mu, \sigma^2 &\sim N(\mu, \sigma^2). \end{aligned}$$

Then the posterior for μ is $\pi(\mu \mid y_{1:n}) \sim N(\mu_1, \sigma_1^2)$ where

$$\mu_1 = \frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum y_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \qquad \sigma_1^2 = \frac{1}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}$$

Do the following:

(a) The tolerance, `epsilon`, is currently set to 0.05. Try running the code with different values for `epsilon`. What conclusions can you draw?

(b) Set `epsilon = 0.005`. The true value for μ is currently set at 0 and σ^2 is 1. What happens to the true posterior as you change these values?

(c) The distance function, `rho`, is currently set to be $\rho(y, x) = |\sum_{i=1}^n y_i - \sum_{i=1}^n x_i|/n$. Try different distance functions to see what happens (e.g. $\rho(y, x) = (\sum_{i=1}^n y_i - \sum_{i=1}^n x_i)^2/n$, KS distance with `R` function `ks.test`). What happens to the ABC-posterior? Why does this happen? How does it impact the tolerances?

(d) Remove the assumption that σ^2 is known. This means you will have to (1) Determine an appropriate prior distribution on σ^2 (e.g. inverse-gamma, uniform), (2) Pick an appropriate summary statistic (*hint: find a sufficient statistic for* σ^2), (3) Set-up a 2-dimensional distance function and a 2-dimensional tolerance to account for the two summary statistics.

3. Problem 3: Sequential Gaussian Example

Using the same model as Problem 2, suppose we have a sample $i = 1, \ldots, n$ of $Y_i \sim N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. Consider the following HBM:

$$\mu \sim N(\mu_0, \sigma_0^2)$$
$$Y_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2).$$

In this problem, we are considering the ABC - Population Monte Carlo algorithm. Do the following:

(a) Run the code and plot the ABC - posteriors at different time steps. How do they compare?

(b) There are currently 10 time steps in the algorithm. Trying changing this to something larger, say 25. Is there a significant improvement in the posterior at $t = 10$ versus $t = 20$? How are you deciding?

(c) As in Problem (2), remove the assumption that σ^2 is known. Use the same set-up as before, but now you have to figure out how to do it sequentially. Use the updating kernel for μ as an example for how to update σ^2.